Bibliometrix: Data Importing and Converting

Massimo Aria

2020-05-25

A common workflow to search and export bibliographic documents

All databases follow a common workflow to search and export bibliographic collections, mainly based on three steps: *1. Write and submit a query*, *2. Refine search results* and *3. Export search results*.

1. Write and submit a query

A query is usually based on a set of terms linked by boolean operators. The search engine will query the db to identify records matching the query.
i.e. TI = (bibliometric AND analysis), the search engine will search for all the records in which the title will contain the words 'bibliometric' and 'analysis' simultaneously.

2. Refine search results

Search results can be refined by applying some filtering criteria for additional fields.
i.e. selecting Document Type = 'Journal Article' AND Language = 'English' AND Timespan = 1990:2020 AND Subject Category = 'Management'

3. Export search results

In this step, the user must choose which metadata he wants to download and the export file format to save the results.

```
To be able to work, *bibliometrix* requires a minimum set of mandatory metadata  (i.e.
Authors' name, Title, Journal title, Affiliation, Publication year, etc.).

Our advice is to always select all the metadata fields to be sure you can perform all the
analyses implemented in bibliometrix.
```

Many databases support a variety of different export formats, some commercial (i.e. EndNote, Mendeley, etc.) and some standard (i.e. html, plaintext, BibTeX, etc.). The choice of a suitable export file will have to consider Table 1.

Supported data sources

Bibliometrix can import bibliographic database files and references manager files, or can download data through APIs.

Bibliometrix supports bibliographic database files from *Web of Science*, *Scopus*, *Dimensions*, *PubMed* and *Cochrane Library*.

In Table 1, we report the complete list of export file formats supported by bibliometrix for each database.

Table 1. *Export file formats supported by bibliometrix*

Source	URL	Format	Extension
Web of Science	https://www.webofknowledge.com/	‘BibTeX’ ‘plaintext’ ’EndNote Desktop	‘.bib’ ‘.txt’ ‘.ciw’
Scopus	https://www.scopus.com/	‘BibTeX’ ‘CSV export’	‘.bib’ ‘.txt’
Dimensions	https://app.dimensions.ai/	‘Bibliometric mapping’ ‘Excel’	‘.csv’ ‘.xlsx’
PubMed	https://pubmed.ncbi.nlm.nih.gov/	‘PubMed export file’	‘.txt’
Cochrane Library	https://www.cochranelibrary.com/	‘plaintext’	‘.txt’

Furthermore, bibliometrix provides support for the APIs (application programming interfaces) of Dimensions, NCBI PubMed and Scopus, using functions from packages dimensionsR (https://github.com/massimoaria/dimensionsR), pubmedR (https://github.com/massimoaria/pubmedR) and rscopus (https://github.com/muschellij2/rscopus), respectively (see Table 2).

Table 2. *API data gathering systems supported by bibliometrix*

Source	API Key request	API Access	Format
Dimensions	Free for scientometric projects https://www.dimensions.ai/scientometric-research/	By an account and password	json
PubMed	Free https://www.ncbi.nlm.nih.gov/account/	Any key (3 requests/s) By a key (10 requests/s)	xml
Scopus	Commercial or Institutional subscription https://dev.elsevier.com/	By a key (10 requests/s)	xml

What types of metadata can be exported

A bibliographic database a very rich set of information about a scientific document that we call '**document metadata**' or '**bibliographic metadata**'.

It is possible to distinguish among five different types of metadata:

- *Document info* (i.e. publication date, journal title, issue, volume, etc.)
- *Authors info* (i.e. authors' name, affiliations, ORCID, etc.)
- *Content info* (i.e. title, abstract, authors' keywords, etc.)
- *Citation info* (i.e. reference lists, number of citations, etc.)
- *Funding info* (i.e. Funding institutes, acknowledgments, etc.)

Web of Science and Scopus allow the user to export the complete set of metadata. This means that it will be possible to perform all analyses implemented in bibliometrix. Some other databases, such as Dimensions, PubMed and Cochrane Library, export just a limited set of metadata types which implies some limitations in the choice of the analyses to carry out.

In Table 3, we show what kind of analyzes can be performed taking into account the metadata set that each database can export.

Table 3. *Type of metadata for each file format*

Source	Format	Exported metadata
Web of Science	◦ 'BibTeX' ◦ 'plaintext' ◦ 'EndNote Desktop'	◦ All ◦ All ◦ All
Scopus	◦ 'BibTeX' ◦ 'csv'	◦ All ◦ All
Dimensions	◦ 'Bibl. mapping' ◦ 'Excel' ◦ 'API'	◦ Document, Authors, Citation info ◦ Document, Authors, Content info ◦ All but with limited reference info
PubMed	◦ 'PubMed export' ◦ 'API'	◦ Document, Authors, Content info ◦ Document, Authors, Content info
Cochrane Library	◦ 'plaintext'	◦ Document, Authors, Content info

How to import a set of export data files

As already stated, *bibliometrix* works with several data formats coming from different sources.

Here, we briefly show how to import and convert data from each bibliographic database.

An export file can be read and converted using the function *convert2df*:

```
**convert2df**(*file*, *dbsource*, *format*)
```

The argument *file* is a character vector containing the export file names. *file* can also contain the name of a JSON/XML object downloaded by Digital Science Dimensions or PubMed APIs (through the packages *dimensionsR* and *pubmedR*.

es. file <- c("file1.txt", "file2.txt", ...)

convert2df creates a bibliographic data frame with cases corresponding to manuscripts and variables to Field Tag in the original export file.

convert2df accepts two additional arguments: *dbsource* and *format*.

The argument *dbsource* indicates from which database the collection has been downloaded.

It can be:

- “isi” or “wos” (for Clarivate Analytics Web of Science database),
- “scopus” (for SCOPUS database),
- “dimensions” (for DS Dimensions database)
- “pubmed” (for PubMed/Medline database),
- “cochrane” (for Cochrane Library database of systematic reviews).

The argument *format* indicates the export file format. It can be one of the formats reported in Table 1.

Each manuscript contains several elements (metadata), such as authors’ names, title, keywords, and other information. All these elements constitute the bibliographic attributes of a document, also called metadata.

Data frame columns are named using the standard Clarivate Analytics WoS Field Tag codify.

The main field tags are:

Field Tag	Description
AU	Authors’ Names
TI	Document Title
SO	Journal Name (or Source)
JI	ISO Source Abbreviation
DT	Document Type
DE	Authors’ Keywords
ID	Keywords associated by SCOPUS or WoS database
AB	Abstract
C1	Authors’ Affiliations
RP	Corresponding Author’s Affiliation
CR	Cited References
TC	Times Cited
PY	Publication Year
SC	Subject Category
UT	Unique Article Identifier
DB	Bibliographic Database

For a complete list of field tags, see
http://www.bibliometrix.org/documents/Field_Tags_bibliometrix.pdf

How to import data from Web of Science

Clarivate Analytics Web of Science (WoS) (http://www.webofknowledge.com) is one of the world's most trusted publisher-independent global citation databases. It was founded by Eugene Garfield, one of the pioneers of bibliometrics. This platform includes many different collections and allows users to export data in several different file formats.

bibliometrix supports three WoS export file formats: BibTeX, plaintext and EndNote Desktop.

To access Web of Science services, a subscription is required. When downloading data from Web of Science, make sure that the Web of Science Core Collection database is selected. Choose the **Export option** followed by **EndNote Desktop** file format, or **Other File Formats** option, and choose either the **Plain Text** or the **BibTeX** file format. Although bibliometrix supports all these file formats, we recommend the use of the plaintext format. When asked which data elements to download, choose the **Full Record and Cited References option**. Downloading cited reference data is necessary for identifying citation, bibliographic coupling, and co-citation links between items.

WoS BibTeX file format

```
library(bibliometrix)

file <- "https://www.bibliometrix.org/datasets/wos_bibtex.bib"

M <- convert2df(file, dbsource = "wos", format = "bibtex")

head(M["TC"])
```

WoS plaintext file format

```
file <- "https://www.bibliometrix.org/datasets/wos_plaintext.txt"

M <- convert2df(file, dbsource = "wos", format = "plaintext")

head(M["TC"])
```

WoS EndNote Desktop file format

```
file <- "https://www.bibliometrix.org/datasets/wos_plaintext.ciw"

M <- convert2df(file, dbsource = "wos", format = "endnote")

head(M["TC"])
```

How to import data from Scopus

Scopus (https://www.scopus.com/) is one of the largest abstract and citation database of peer-reviewed literature: scientific journals, books, and conference proceedings. Like WoS, Scopus allows to export data in several different file formats.

bibliometrix supports two Scopus export file formats: BibTeX and csv (comma-separated values).

TO access Scopus services, a subscription is required. To download data from Scopus, choose the **CSV** or the **BibTeX export** option. (Do not choose the Download option!) Make sure that the data is downloaded in a '.csv' or '.bib' file and that all data elements are included. Although bibliometrix supports both file formats, we recommend the use of the CSV format.

Scopus BibTeX file format

```
library(bibliometrix)

file <- "https://www.bibliometrix.org/datasets/scopus_csv.csv"

M <- convert2df(file, dbsource = "scopus", format = "csv")

head(M["TC"])
```

Scopus csv file format

```
library(bibliometrix)

file <- "https://www.bibliometrix.org/datasets/scopus_bibtex.bib"

M <- convert2df(file, dbsource = "scopus", format = "bibtex")

head(M["TC"])
```

How to import data from Dimensions

Dimensions is s comprehensive database by Digital Science that doesn't impose limitations on users. Like WoS, Scopus lets to export data in several different file formats. Dimensions is the only database that links publications and citations with grants, patents, clinical trials, datasets, and policy papers to deliver a more holistic view of the research landscape. Differently from WoS and Scopus, Dimensions allows exporting data in only two different file formats: excel and csv.

These two formats are not interchangeable but contain different types of metadata. In particular, Excel file contains metadata about document contents such as Abstract, Keywords, etc.. On the contrary, the csv file contains reference lists and affiliation names that can be used in citation and co-citation analyzes.

bibliometrix supports both the two Dimensions export file formats: excel and csv (comma-separated values).

The free version of Dimensions, for which no subscription is needed, can be used. A user account is required. To download data from Dimensions, choose the Save / Export option, followed by the **Export for bibliometric mapping or Excel option**.

Dimensions Excel file format

```
file <- "https://www.bibliometrix.org/datasets/dimensions_excel.xlsx"

M <- convert2df(file, dbsource = "dimensions", format = "excel")

head(M["TC"])
```

Dimensions csv file format

```
file <- "https://www.bibliometrix.org/datasets/dimensions_csv.csv"

M <- convert2df(file, dbsource = "dimensions", format = "csv")

head(M["TC"])
```

How to import data from PubMed

PubMed comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubMed allows user to export data in many different file formats but the most suitable for science mapping analysis is the "pubmed txt" format. However, Any of these file formats contain metadata about citations and reference lists. This means that is not possible to perform citation and co-citation analysis using PubMed data.

bibliometrix supports only the "pubmed txt" file format.

PubMed can be accessed at https://pubmed.ncbi.nlm.nih.gov/. To download data from PubMed, choose the **Save option**, choose **All results** as content selection, and choose **PubMed** as the file format. Data downloaded from PubMed cannot be used for identifying citation, bibliographic coupling, and co-citation links between items.

PubMed txt file format

```
file <- "https://www.bibliometrix.org/datasets/pubmed_txt.txt"

M <- convert2df(file, dbsource = "pubmed", format = "pubmed")

head(M["TC"])
```

How to import data from Cochrane Library

The *Cochrane Database of Systematic Reviews (CDSR)*, owned by Cochrane Library, is a leading journal and database for systematic reviews in health care. CDSR includes Cochrane Reviews (systematic reviews) and protocols for Cochrane Reviews as well as editorials and supplements.

As for PubMed, CDSR allows to export many file formats but any of these contains metadata about citations and reference lists.

bibliometrix supports the "plaintext" file format.

CDSR can be accessed at https://www.cochranelibrary.com/search. To download data from CDSRed, choose the **Export selected citation(s) option**, and choose **Plain Text** as the file format. Data downloaded from CDSR cannot be used for identifying citation, bibliographic coupling, and co-citation links between items.

Cochrane plaintext file format

```
file <- "https://www.bibliometrix.org/datasets/cochrane_plaintext.txt"

M <- convert2df(file, dbsource = "cochrane", format = "plaintext")

head(M["TC"])
```

How to download bibliographic data using APIs

bibliometrix can download data through an API. The APIs supported by bibliometrix are listed in Table 2. The use of APIs requires an internet connection. bibliometrix will download data for all documents that match the specified search criteria defined by a query.

Dimensions API needs an account to obtain a valid token to query the database. The account can be obtain for free for scientometric reseaech project asking for it at https://www.dimensions.ai/scientometric-research/. Scopus API needs an API Key to work. Full Scopus APIs access is only granted to users with Scopus subscription (i.e. an academic subscription). These users can be request an API key at https://dev.elsevier.com/. FInally, by default, the access to PubMed API is free and does not necessarily require an API key. In this case, PubMed limits users to making only 3 requests per second. Users who register for an API key are able to make up to ten requests per second.

Obtaing a key is very simple, you just need to register for “my ncbi account” (https://www.ncbi.nlm.nih.gov/account/) then click on a button in the “account settings page” (https://www.ncbi.nlm.nih.gov/account/settings/).

At the moment, an important limitation of the Dimensions and PubMed APIs is that data downloaded through this API cannot be used for identifying citation, bibliographic coupling, and co-citation links between items.

How to import data using Dimensions API

The *Dimensions API* provides access to Dimensions data directly, and makes it possible to retrieve results to precise and complex queries. These are performed using the *dimensionsR R-package* functions which implement the Dimensions Search Language (DSL), a custom query language created by Dimensions.

The user can choose to write a valid query using that language or, in alternaative, using the function *dsQueryBuild*.

For a more detailed focus on the use of the dimensionsR package, please see the package vignette at https://cran.r-project.org/package=dimensionsR/vignettes/A_Brief_Example.html

How to import data using PubMed API

The *PubMed API* provides access to PubMed data directly, submitting a query written following the pubmed query language. The PubMed data can be gathered using the *pubmedR R-package* functions.

For a more detailed focus on the use of the pubmedR package, please see the package vignette at https://cran.r-project.org/package=pubmedR/vignettes/A_Brief_Example.html